Hardware Recommendations for Data Science

Our hardware recommendations for data science and analysis workstations below were provided by Dr. Kinghorn. These follow some standard patterns, but keep in mind that your specific workflow may have unique requirements.

Browse Our Recommended Systems

Puget Labs Certified

These hardware configurations have been developed and verified through frequent testing by our Labs team. Click here for more details.

Hardware Recommendations

Suggested Systems

Intel Xeon Workstation

AMD Threadripper PRO Workstation

Data Science System Requirements

Quickly Jump To: Processor (CPU) • Video Card (GPU) • Memory (RAM) • Storage (Drives)

Data Science / Data Analysis is coupled with methods from machine learning, so there are some similarities here with our Hardware Recommendations for ML/AI. However, data analysis, preparation, munging, cleaning, visualization, etc does present unique challenges for system configuration. Extract, Transform, and Load (ETL) and Exploratory Data Analysis (EDA) are critical components of machine learning projects, as well as being indispensable parts of business processes and forecasting.

The “best” hardware will follow some standard patterns, but your specific application may have unique optimal requirements. The Q&A discussion below, with answers provided by Dr. Donald Kinghorn, will be mostly generalities based on typical workflows. We also recommend that you visit his HPC blog for more info.

Processor (CPU)

In data science there is a significant amount of effort with movement and transformation of large data sets. The CPU, with its ability to access large amounts of memory, may dominate workflows in contrast to GPU compute in ML/DL. Multi-core parallelism will depend on the task, but parallelism in data processing is often very good.

What CPU is best for data science?

The two recommended CPU platforms are Intel's Xeon® 600 and AMD's Threadripper™ PRO. Both of these offer high core counts, excellent memory performance & capacity, and large numbers of PCIe lanes. Specifically, the 32-core versions of either of these are recommended for their utilization and balanced memory performance.

Do more CPU cores make data science workflows faster?

The number of cores chosen will depend on the expected load and parallelism of tasks in your workflow. Larger numbers of cores may also allow for multiple simultaneous processes. An easy recommendation is for 32 cores with either of the Intel or AMD platforms mentioned above. The 96- or 64-core TR PRO may be ideal if you have highly data parallel tasks with a significant amount of time spent in computation, but scaling may not be as efficient as with the 32-core if memory access is a limiting factor. In any case, a 16-core processor would probably be considered minimal.

Does data science work better with Intel or AMD CPUs?

It is mostly a matter of preference. However, the Intel Xeon platform would be recommended if your workflow could benefit from some of the tools in the Intel oneAPI AI Analytics Toolkit, such as the Pandas alternative Modin which is optimized for Intel, or Advanced Matrix Extenions (AMX).

Looking for a Data Science Workstation?

We build computers tailor-made for your workflow.

CONFIGURE A WORKSTATION

Don't know where to start?
We can help!

Get in touch with our technical consultants today.

Talk to an Expert

Video Card (GPU)

Since the mid-2010s, GPU acceleration has been the driving force enabling rapid advancements in machine learning and AI research. NVIDIA has had a massive impact in this field. For data science, the GPU may offer significant performance over the CPU for some tasks. However, GPUs may be limited by memory capacity and appropriate applications for data tasks outside of model training.

What type of GPU (video card) is best for data science?

NVIDIA dominates for GPU compute acceleration, and is unquestionably the standard. Their GPUs will be the most supported and easiest to work with. NVIDIA also provides an excellent data-handling application suite called RAPIDS. **The NVIDIA RAPIDS tools may provide significant workflow throughput.**

How much VRAM (video memory) does data science need?

This is dependent on the "feature space" of your data. Memory capacity on GPUs is limited compared to the main system memory utilized by CPUs, and applications may be constrained by this. This is why it's common for a data scientist to be tasked with "data and feature reduction" prior to model training. That is often 80+% of the hard work for ML/AI projects. For some jobs, GPU memory may be a limiting factor even when there is a GPU-accelerated tool available for the data work. For larger data problems, the 96GB available on the NVIDIA RTX PRO™ 6000 Blackwell may be necessary – and even that may not be enough for jobs that require all data to be resident on the device. Data movement can be a bottleneck because GPUs have such highly performant compute capabilities that they may be left idle a large percent of the time while waiting for memory to move around.

Will multiple GPUs improve performance in data science workflows?

For data analysis jobs that can take advantage of GPUs, having more than one may increase workflow. If you will be doing ML/AI jobs then multi-GPU can be beneficial since many frameworks provide for this. For data-oriented tasks, multi-GPU may have an advantage simply by providing more available memory to facilitate task parallelism. Not all workflows utilize the GPU well, though, as discussed previously.

Do I need NVLink when using multiple GPUs for data science?

NVIDIA's NVLink provides a direct, high-performance communication bridge between GPUs. It has become less important as PCIe speeds have gotten faster, and as such has been removed from most NVIDIA graphics cards. The high-performance H200 NVL is one of the few models that still supports NVLink, but it requires a specialized bridge.

Looking for a Data Science Workstation?

We build computers tailor-made for your workflow.

CONFIGURE A WORKSTATION

Don't know where to start?
We can help!

Get in touch with our technical consultants today.

Talk to an Expert

Memory (RAM)

CPU Memory capacity may be the limiting factor for some data analysis tasks. This is because an entire large data set may need to be resident in memory (in-core). There are methods and tools for "out-of-core" data analysis, but this can slow performance.

How much RAM does data science need?

It is often necessary, or at least desirable, to be able to pull a full data set into memory for processing and statistical work. That could mean BIG memory requirements, as much as 1-2 TB of system memory for the CPU to access.

Storage (Hard Drives)

Storage requirements are similar to CPU memory requirements. Your data and projects will dictate requirements.

What storage configuration works best for data science?

It's recommended to use fast NVMe storage whenever possible since data streaming can become a bottleneck when data is too large to fit in system memory. Staging job runs from NVMe can reduce job run slow ups. NVMe and SATA solid-state drives are available up to 8TB capacity, with NVMe drives being much faster and generally preferred. Platter drives can be used for archival storage and for very large data sets, but should not be used for active working space. They are available in capacities exceeding 20TB now.

Additionally, all of the above drive types can be configured in RAID arrays. This does add complexity to the system configuration and may use up slots on the motherboard which would otherwise support additional GPUs – but can allow for storage space in the 10 to 100s of terrabytes.

Should I use network attached storage for data science?

Network-attached storage is another consideration. It's become more common for workstation motherboards to have 10Gb Ethernet ports, allowing for network storage connections with reasonably good performance without the need for more specialized networking add-ons. Rackmount workstations and servers can have even faster network connections, often using more advanced cabling than simple RJ45, making options like software-defined storage appealing.

Looking for a Data Science workstation?

We build computers that are tailor-made for your workflow.

Configure a System

Don't know where to start? We can help!

Get in touch with one of our technical consultants today.

Talk to an Expert

Related Content

- Standing Up AI Development Quickly for Supercomputing 2025
- Exploring Hybrid CPU/GPU LLM Inference
- What's the deal with NPUs?
- AMD Zen4 Threadripper PRO vs Intel Xeon-w9 For Science and Engineering

View All Related Content

Latest Content

- The Shovel is Real: Why AI is the New Layer of Compute
- HOW TO: Install and Configure Puget Systems Docker App Packs
- From Unboxing to Inference: Introducing the Puget Systems Docker App Packs
- Standing Up AI Development Quickly for Supercomputing 2025

View All